(特許協力条約に基づいて公開された国際出願)

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2003年7月24日 (24.07.2003)

PCT

(10) 国際公開番号
WO 03/060878 A1

(51) 国際特許分類[7]: G10L 15/06, 15/18

(21) 国際出願番号: PCT/JP02/13053

(22) 国際出願日: 2002年12月13日 (13.12.2002)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願2002-007283 2002年1月16日 (16.01.2002) JP

(71) 出願人 (米国を除く全ての指定国について): シャープ株式会社 (SHARP KABUSHIKI KAISHA) [JP/JP]; 〒545-8522 大阪府 大阪市 阿倍野区長池町２２番２２号 Osaka (JP).

(72) 発明者; および
(75) 発明者/出願人 (米国についてのみ): 鶴田 彰 (TSURUTA,Akira) [JP/JP]; 〒639-1135 奈良県 大和郡山市 天井町１６７－３ Nara (JP).

(74) 代理人: 青山 葆, 外(AOYAMA,Tamotsu et al.); 〒540-0001 大阪府 大阪市 中央区城見１丁目３番７号 ＩＭＰビル 青山特許事務所 Osaka (JP).

(81) 指定国 (国内): CN, KR, US.

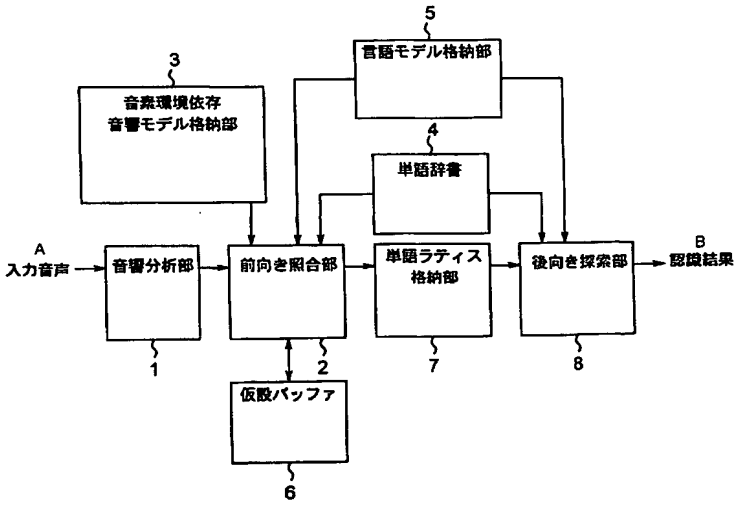
(84) 指定国 (広域): ヨーロッパ特許 (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SI, SK, TR).

添付公開書類:
— 国際調査報告書

[続葉有]

(54) Title: CONTINUOUS SPEECH RECOGNITION APPARATUS, CONTINUOUS SPEECH RECOGNITION METHOD, CONTINUOUS SPEECH RECOGNITION PROGRAM, AND PROGRAM RECORDING MEDIUM

(54) 発明の名称: 連続音声認識装置および連続音声認識方法、連続音声認識プログラム、並びに、プログラム記録媒体



3...PHONEME ENVIRONMENT DEPENDENT ACOUSTIC MODEL STORAGE UNIT
5...LANGUAGE MODEL STORAGE UNIT
4...WORD DICTIONARY
A...INPUT SPEECH
1...ACOUSTIC ANALYSIS UNIT
2...FORWARD MATCHING UNIT
7...WORD LATTICE STORAGE UNIT
8...BACKWARD SEARCH UNIT
B...RECOGNITION RESULT
6...TEMPORARY BUFFER

(57) Abstract: It is possible to assure accuracy even at the word boundary by using the phoneme environment dependent acoustic model and suppress increase of the processing amount even when recognizing a continuous speech of large vocabulary. A phoneme environment dependent acoustic model storage unit (3) contains a phoneme state tree, i.e., a tree structure of state series of the preceding phoneme state, central phoneme state, and subsequent phoneme state while collecting a triphone model having the same preceding phoneme and the central phoneme. Accordingly, in order to spread a phoneme hypothesis by referencing the phoneme state tree, the language model

[続葉有]

WO 03/060878 A1

stored in the language model storage unit (5), and the word dictionary (4) by the forward matching unit (2), what is necessary is only to spread one phoneme hypothesis regardless of the head phoneme of the subsequent word. Thus it is possible to easily spread a hypothesis regardless of in-word or word-boundary state. Moreover, it is possible to significantly reduce the matching amount when performing matching with the characteristic parameter series from an acoustic analysis unit (1).

(57) 要約:

単語境界にも音素環境依存音響モデルを用いて精度を確保しつつ大語彙の連続音声認識時にも処理量の増大を抑える。音素環境依存音響モデル格納部（３）には、先行音素および中心音素が同じトライフォンモデルをまとめて先行音素の状態と中心音素の状態と後続音素の状態との状態系列を木構造化した音素状態木を格納している。したがって、前向き照合部（２）によって、上記音素状態木，言語モデル格納部（５）に格納された言語モデルおよび単語辞書（４）を参照して音素仮説を展開する際には、次に続く単語の先頭音素に関係無く１つの音素仮説を展開すればよく、単語内および単語境界に関係なく仮説の展開が容易になる。また、音響分析部（１）からの特徴パラメータ系列との照合を行う際における照合処理量を大幅に削減できる。

明　細　書

連続音声認識装置および連続音声認識方法、連続音声認識プログラム、並びに、プログラム記録媒体

技術分野

この発明は、音素環境依存音響モデルを用いて高精度に認識を行う連続音声認識装置および連続音声認識方法、連続音声認識プログラム、並びに、連続音声認識プログラムを記録したプログラム記録媒体に関する。

背景技術

一般に、大語彙連続音声認識で用いる認識単位としては、認識対象語彙の変更や大語彙への拡張が容易であることから、音節や音素等の単語より小さいサブワードと呼ばれる認識単位が用いられることが多い。さらに、調音結合等の影響を考慮するためには、前後の環境(コンテキスト)に依存したモデルが有効であることが知られている。例えば、前後一つずつの音素に依存したトライフォンモデルと呼ばれる音素モデルが広く使用されている。

また、連続的に発声された音声を認識する連続音声認識方法の一つとして、語彙中の各単語をサブワードのネットワークや木構造等で記述したサブワード表記辞書と、単語の接続の制約を記述した文法または統計的言語モデルの情報とに従って、単語を連結して認識結果を得る方法がある。

これらのサブワードを認識単位とした連続音声認識技術については、例えば、刊行物「音声認識の基礎(下)」古井貞熙監訳に詳しく説明されている。

上述したごとく、環境に依存したサブワードを用いて連続音声認識を行う場合には、単語内だけではなく単語間においても音素環境依存型の音響モデルを用いた方が、認識精度がよいことが知られている。しかしながら、単語の始終端に用いる音響モデルは前後に接続する単語に依存するため、音素環境に依存しない音響モデルを用いる場合に比べて、処理が複雑になると共に処理量が大幅に増えてしまう。

以下、単語辞書と言語モデルと音素環境依存音響モデルを参照して、単語履歴毎に木を動的に生成する方法について、具体的に説明する。

例えば、「朝の天気…」という発声に対して、「朝(a;s;a)」という単語の最後の音素/a/を考える場合、図３に示す単語辞書の情報から得られる単語「朝日(a;s;a;h;i)」における３番目の音素/a/とその前後に続く音素とから成るトライフォン“s;a;h”と、図４に示す言語モデルの情報から得られる単語「の(n;o)」とその前に続く単語「朝(a;s;a)」との連鎖「朝の(a;s;a;n;o)」における３番目の音素/a/とその前後に続く音素とから成るトライフォン“s;a;n”とについて、仮説を展開する必要がある。この例の場合は２つの仮説を展開するだけでよいが、より複雑な文法や統計的言語モデルを用いる場合には、単語の終端で多くの単語につながる可能性がある。そして、その場合には、それらの先頭の音素に依存して、例えば図２Ｂに示すような先行音素と中心音素と後続音素からなるトライフォンの状態系列を用いて、図５Ｂに示すように多くの仮説を展開する必要がある。

この問題に対し、単語内には音素環境依存の音響モデルを用いる一方、単語境界では環境に依存しない音響モデルを使用する連続音声認識方式が、特開平５－２２４６９２号公報に開示されている。この連続音声認識方式によれば、単語間での処理量の増大を抑えることができる。また、認識対象語彙中の各単語について、前後の単語に依存せずに決まる音響モデル系列を認識単語として記述した認識単語辞書と、単語境界において前後の単語に依存して記述した単語間単語辞書とを用いて照合する連続音声認識方式が、特開平１１－４５０９７号公報に開示されている。この連続音声認識方式によれば、単語境界に音素環境依存の音響モデルを用いても処理量の増大を抑えることができるのである。

しかしながら、上記従来の連続音声認識方式においては、以下のような問題がある。すなわち、特開平５－２２４６９２号公報に開示された連続音声認識方式においては、単語内には音素環境依存の音響モデルを用い、単語境界では環境に依存しない音響モデルを用いている。したがって、単語境界での処理量の増大を抑えることはができるが、その一方において、単語境界に用いる音響モデルの精度が低いために、特に大語彙の連続音声認識の場合には認識性能の低下を招く恐れがある。

これに対して、特開平１１－４５０９７号公報に開示された連続音声認識方式においては、前後の単語に依存せずに決まる音響モデル系列を認識単語として記述した認識単語辞書と、単語境界において前後の単語に依存して記述した単語間単語辞書を用いて照合を行うようにしている。したがって、単語境界にも音素環境依存の音響モデルを用いることによって精度を確保しながら、大語彙の場合でも単語境界での処理量の増大を抑えることができるのである。しかしながら、一般に、単語のスコアや境界はそれ以前の単語の影響を受けるので、複数の認識単語が単語間単語を共有すると、図９Ａに示すように認識単語“k;o;k”および“s;o;k”と単語間単語“o”との境界の履歴が考慮されないので、図９Ｂに示すように単語の境界履歴を考慮した場合に比して、性能の低下を招く恐れがある。また、例えば助詞の“を(/ｏ/と発声)”等のように、認識単語辞書と単語間単語辞書とに分割することができない単語についは開示されてはいない。

発明の開示

そこで、この発明の目的は、単語境界にも音素環境依存音響モデルを用いて精度を確保しつつ、大語彙の連続音声認識時にも単語境界での処理量の増大を抑えることができる連続音声認識装置および連続音声認識方法、連続音声認識プログラム、並びに、連続音声認識プログラムを記録したプログラム記録媒体を提供することにある。

上記目的を達成するため、この発明は、隣接するサブワードに依存して決定されるサブワードを認識単位とすると共に,サブワード環境に依存する環境依存音響モデルを用いて,連続的に発声された入力音声を認識する連続音声認識装置であって、入力音声を分析して特徴パラメータの時系列を得る音響分析部と、語彙中の各単語が,サブワードのネットワークあるいはサブワードの木構造として格納された単語辞書と、単語間の接続情報を表す言語モデルが格納された言語モデル格納部と、上記環境依存音響モデルが,当該環境依存音響モデルの状態系列のうち,複数のサブワードモデルの状態系列をまとめて木構造化して成るサブワード状態木として格納されている環境依存音響モデル格納部と、上記環境依存音響モデルであるサブワード状態木,上記単語辞書および言語モデルを参照して上記

サブワードの仮説を展開すると共に，上記特徴パラメータの時系列と上記展開された仮説との照合を行い，単語の終端に該当する仮説に関する単語，累積スコア及び始端開始フレームを含む単語情報を単語ラティスとして出力する照合部と、上記単語ラティスに対する探索を行って認識結果を生成する探索部を備えたことを特徴としている。

上記構成によれば、サブワード環境に依存する環境依存音響モデルを木構造化したサブワード状態木，単語辞書および言語モデルを参照して、サブワードの仮説を展開するようにしている。したがって、次に続く単語の先頭サブワードに関係無く１つの仮説を展開すればよく、全仮説における状態の総数を削減することができる。すなわち、仮説の展開処理量を大幅に削減でき、単語内および単語境界に関係なく、仮説の展開が容易になるのである。さらに、照合部によって、上記音響分析部からの特徴パラメータ系列と上記展開された仮説との照合を行う際における照合処理量が大幅に削減される。

また、１実施例では、上記発明の連続音声認識装置において、上記環境依存音響モデル格納部に格納されている環境依存音響モデルは、中心サブワードが前後のサブワードに依存する環境依存音響モデルのうち、先行サブワードおよび中心サブワードが同じサブワードモデルの状態系列を木構造化したサブワード状態木である。

この実施例によれば、先行サブワードおよび中心サブワードが同じサブワードモデルの状態系列を木構造化したサブワード状態木を用いて、上記仮説を展開している。したがって、次の仮説を展開する場合には、終端仮説における中心サブワードのみに注目して対応する先行サブワードを有するサブワード状態木を展開すればよい。つまり、後続サブワードが複数あってもより少ない仮説を展開すればよく、仮説の展開が容易である。

また、１実施例では、上記発明の連続音声認識装置において、上記環境依存音響モデルは、複数のサブワードモデルで状態を共有している状態共有モデルである。

この実施例によれば、複数のサブワードモデルによって状態を共有することによって、木構造化した際に共有している状態を一つにまとめることができ、ノー

ド数を削減することができる。したがって、上記照合部による照合時における処理量が大幅に削減される。

また、１実施例では、上記発明の連続音声認識装置において、上記照合部は、上記サブワード状態木を参照して仮説を展開する際に、上記単語辞書および言語モデルから得られる接続可能なサブワード情報を用いて、上記仮説であるサブワード状態木を構成する状態のうち、互いに接続可能な状態にフラグを付すようになっている。

この実施例によれば、上記展開された仮説を構成するサブワード状態木の状態のうち、互いに接続可能な状態のみにフラグを付けるようにしたので、上記照合の際にビタビ計算を行う必要がある状態が限定されて、照合処理量が更に削減される。

また、１実施例では、上記発明の連続音声認識装置において、上記照合部は、上記照合を行う際に、上記特徴パラメータの時系列に基づいて上記展開された仮説のスコアを算出すると共に、このスコアの閾値あるいは仮説数を含む基準に従って上記仮説の枝刈りを行うようになっている。

この実施例によれば、上記照合時に仮説の枝刈りを行うので、単語となる可能性が低い仮説が削除されて、以後の照合処理量が大幅に削減される。

また、この発明は、隣接するサブワードに依存して決定されるサブワードを認識単位とすると共に，サブワード環境に依存する環境依存音響モデルを用いて，連続的に発声された入力音声を認識する連続音声認識方法であって、音響分析部によって，上記入力音声を分析して特徴パラメータの時系列を得、照合部によって，上記環境依存音響モデルの状態系列を木構造化して成るサブワード状態木，語彙中の各単語がサブワードのネットワークあるいはサブワードの木構造として記述された上記単語辞書，および，単語間の接続情報を表す言語モデルを参照して，上記サブワードの仮説を展開すると共に，上記特徴パラメータの時系列と上記展開された仮説との照合を行って，単語の終端に該当する仮説に関する単語，累積スコアおよび始端開始フレームを含む単語情報を単語ラティスとして生成し、探索部によって，上記単語ラティスに対する探索を行って認識結果を生成することを特徴としている。

上記構成によれば、上記発明の連続音声認識装置の場合と同様に、環境依存音響モデルを木構造化したサブワード状態木を参照して仮説を展開するので、次に続く単語の先頭サブワードに関係無く１つの仮説を展開すればよく、単語内および単語境界に関係なく仮説の展開が容易になるのである。さらに、特徴パラメータ系列と上記展開された仮説との照合を行う際における照合処理量が大幅に削減される。

また、この発明の連続音声認識プログラムは、コンピュータを、上記発明の連続音声認識装置における音響分析部, 単語辞書, 言語モデル格納部, 環境依存音響モデル格納部, 照合部および探索部として機能させることを特徴としている。

上記構成によれば、上記発明の連続音声認識装置の場合と同様に、次に続く単語の先頭サブワードに関係無く１つの仮説を展開すればよく、単語内および単語境界に関係なく仮説の展開が容易になる。さらに、特徴パラメータ系列と上記展開された仮説との照合を行う際における照合処理量が大幅に削減される。

また、この発明のプログラム記録媒体は、上記発明の連続音声認識プログラムが記録されたことを特徴としている。

上記構成によれば、上記発明の連続音声認識装置の場合と同様に、次に続く単語の先頭サブワードに関係無く１つの仮説を展開すればよく、単語内および単語境界に関係なく仮説の展開が容易になる。さらに、特徴パラメータ系列と上記展開された仮説との照合を行う際における照合処理量が大幅に削減される。

図面の簡単な説明

図１は、この発明の連続音声認識装置におけるブロック図である。

図２Ａ、図２Ｂは、音素環境依存音響モデルの説明図である。

図３は、図１における単語辞書の説明図である。

図４は、言語モデルの説明図である。

図５Ａ、図５Ｂは、図１における前向き照合部による仮説の展開の説明図である。

図６は、上記前向き照合部によって実行される前向き照合処理動作のフローチャートである。

図７Ａ、図７Ｂは、上記前向き照合部による仮説の照合および仮説の枝刈りの説明図である。

図８は、音素仮説の音素状態木における必要な状態のみにフラグを付す場合の説明図である。

図９は、認識単語と単語間単語との境界の履歴が考慮されない場合と考慮された場合との比較図である。

発明を実施するための最良の形態

以下、この発明を図示の実施の形態により詳細に説明する。図１は、本実施の形態の連続音声認識装置におけるブロック図である。この連続音声認識装置は、音響分析部１，前向き照合部２，音素環境依存音響モデル格納部３，単語辞書４，言語モデル格納部５，仮説バッファ６，単語ラティス格納部７および後向き探索部８で構成される。

図１において、入力音声は、音響分析部１によって、特徴パラメータの系列に変換されて前向き照合部２に出力される。前向き照合部２では、音素環境依存音響モデル格納部３に格納された音素環境依存音響モデル，言語モデル格納部５に格納された言語モデルおよび単語辞書４を参照して、仮説バッファ６上に音素仮説を展開する。そして、上記音素環境依存音響モデルを用いて、上記展開された音素仮説と特徴パラメータ系列との照合をフレーム同期ビタビビームサーチによって行い、単語ラティスを生成して単語ラティス格納部７に格納する。

上記音素環境依存音響モデルとしては、トライフォンモデルと呼ばれる前後一つずつの音素環境を考慮した隠れマルコフモデル(HMM)を用いている。すなわち、上記サブワードモデルは音素モデルである。但し、従来においては図２Ｂに示すように中心音素の前後１つずつの先行音素と後続音素とを考慮したトライフォンモデルを３状態の状態系列(状態番号列)で表現していたものを、本実施の形態においては、図２Ａに示すように、先行音素と中心音素とが同じトライフォンモデルの状態系列をまとめて木構造(以下、音素状態木という)化している。図２Ｂに示すように、複数のトライフォンモデルで状態を共有している状態共有モデルは、状態系列を木構造化して音素状態木を作成することによって状態数を削減

することができ、計算量の削減を行うことができるのである。

上記単語辞書４としては、認識対象語彙の各単語について、その単語の読みを音素系列で表記し、図３に示すように、上記音素系列を木構造化したものを用いる。言語モデル格納部５には、例えば、図４に示すように、文法によって設定された単語間の接続情報が言語モデルとして格納されている。尚、本実施の形態においては、単語の読みを表わす音素系列を木構造化したものを単語辞書４としているが、ネットワーク化したものでも差し支えない。また、言語モデルとして文法モデルを用いたが、統計的言語モデルを用いても差し支えない。

上記仮説バッファ６上には、上述したように、上記前向き照合部２によって、音素環境依存音響モデル格納部３，単語辞書４および言語モデル格納部５が参照されて、図５Ａに示すような音素仮説が順次展開される。後向き探索部８は、言語モデル格納部５に格納された言語モデルおよび単語辞書４を参照しながら、単語ラティス格納部７に格納されている単語ラティスを、例えばＡ＊アルゴリズムを用いて探索することによって、入力音声に対する認識結果を得るようになっている。

以下、上記前向き照合部２によって、上記音素環境依存音響モデル格納部３，単語辞書４および言語モデル格納部５を参照して、仮説バッファ６上に仮説を展開して単語ラティスを生成する方法について、図６に示す前向き照合処理動作フローチャートに従って説明する。

ステップＳ1で、先ず照合を始める前に仮説バッファ６の初期化を行う。そして、無音から各単語の始端に続く“-;-;＊”なる音素状態木が初期仮説として仮説バッファ６にセットされる。ステップＳ2で、上記音素環境依存音響モデルが用いられて、処理対象のフレームにおける特徴パラメータと仮説バッファ６内にある図７Ａに示すような音素仮説との照合が行われ、各音素仮説のスコアが計算される。ステップＳ3で、図７Ｂに示すように、上記スコアの閾値あるいは仮説数等に基づいて、仮説１及び仮説４のように音素仮説の枝刈りが行われる。こうして、音素仮説の不必要な増大が防止される。ステップＳ4で、仮説バッファ６内に残っている音素仮説のうち単語終端がアクティブなものについて、単語，累積スコアおよび始端開始フレーム等の単語情報が単語ラティス格納部７に保存さ

れる。こうして、単語ラティスが生成されて保存される。ステップＳ5で、図７Ｂに示される仮説５および仮説６のように、音素環境依存音響モデル格納部３，単語辞書４および言語モデル格納部５の情報が参照されて、仮説バッファ６内に残っている音素仮説が伸ばされる。ステップＳ6で、当該処理対象フレームは最終フレームであるか否かが判別される。その結果、最終フレームである場合には前向き照合処理動作を終了する。一方、最終フレームでない場合には上記ステップＳ2に戻って、次のフレームの処理に移行する。そして、以後、上記ステップＳ2～ステップＳ6までが繰り返され、上記ステップＳ6において最終フレームであると判別されると前向き照合処理動作を終了する。

以下、上記前向き照合処理動作の際に、先行音素および中心音素が同じであるトライフォンモデルの状態系列が木構造化された音素状態木を用いる場合の効果について説明する。

例えば、「朝の天気…」という発声に対して、「朝(a;s;a)」という単語の最後の音素/ａ/を考える場合に、図３に示す単語辞書４の情報から得られた単語「朝日(a;s;a;h;i)」における３番目の音素/ａ/とその前後に続く音素とから成るトライフォン“s;a;h”と、図４に示す言語モデルの情報から得られた単語「の(n;o)」とその前に続く単語「朝(a;s;a)」との連鎖「朝の(a;s;a;n;o)」における３番目の音素/ａ/とその前後に続く音素とから成るトライフォン“s;a;n”とについて、音素仮説を展開することが可能である。この場合には２つの音素仮説を展開するだけでよいが、より複雑な文法や統計的言語モデルを参照した場合には単語の終端で多くの次の単語につながる可能性があり、図５Ｂに示すように、次の単語の先頭音素に応じて多数の音素仮説を展開することになる。これに対して、本実施の形態のように音素状態木の音素仮説を展開する場合には、次の単語の先頭音素に関係なく図２Ａに示すような音素状態木“s;a;＊”を、図５Ａに示すように１つ展開するだけでよいのである。尚、図５Ａにおいては、音素状態木のシンボルとして「木」を模した三角形を当てている。

ところで、図５Ｂに示すように、個々の音素について仮説を展開する場合には、次に続く単語の先頭音素の種類を全２７とした場合、新たに展開される音素仮説の数は２７となり、その新たに展開される全音素仮説における状態の総数は８１

(＝２７×３)となる。

これに対して、図５Ａに示すように、上記音素状態木を用いて音素仮説を展開することによって、新たに展開される音素仮説の数は１となり、その状態の総数は２９(１＋７＋２１)に削減することができる。したがって、仮説の展開処理および照合処理の処理量を大幅に削減できるのである。

また、上記言語モデルに文法を用いる場合、単語辞書４および言語モデルによって後続の音素が限定されることが多い。そこで、図８に示すように、音素状態木“s;a;＊”の各状態のうち、単語辞書４に基づく音素列“s;a;h”および言語モデルに基づく音素列“s;a;n”に必要な状態のみにフラグ(図８中においては楕円印)を付すことによって、照合の全状態数を、音素状態木“s;a;＊”の総ての状態数２９に比して状態数５に削減できる。したがって、照合の処理量を更に削減できるのである。

以上のごとく、本実施の形態においては、音素環境依存音響モデル格納部３には、先行音素および中心音素が同じトライフォンモデルの状態系列をまとめて木構造化した音素状態木を格納している。その結果、複数のトライフォンモデルで状態を共有している状態共有モデルの場合には、木構造化した際に共有されている状態を一つにまとめることができ、ノード数を削減することができる。したがって、個々の音素について仮説を展開する場合に上記音素状態木を音素仮説として用いることによって、次に続く単語の先頭音素に関係無く１つの音素仮説を展開すればよいことになる。したがって、次に続く単語の先頭音素の種類を全２７と仮定した場合、従来は、新たに２７個の音素仮説が展開されるために全音素仮説における状態の総数は８１となる。これに対して、本実施の形態においては、新たに展開される音素仮説は１個であるために全音素仮説における状態の総数を２９に削減することができるのである。

すなわち、本実施の形態によれば、上記前向き照合部２によって、音素環境依存音響モデル格納部３に格納された音素環境依存音響モデル，言語モデル格納部５に格納された言語モデルおよび単語辞書４を参照して音素仮説を展開する際における音素仮説の展開処理量を大幅に削減できる。したがって、単語内および単語境界に関係なく、仮説の展開が容易になる。また、前向き照合部２によって、

上記音素環境依存音響モデルを用いて、音響分析部１からの特徴パラメータ系列と上記展開された音素仮説とのフレーム同期ビタビビームサーチによる照合を行う際における照合処理量を大幅に削減できるのである。

また、その際に、上記前向き照合部２は、上記音素仮説との照合を行う際に、各音素仮説のスコアを計算し、スコアの閾値あるいは仮説数の閾値に基づいて音素仮説の枝刈りを行うようにしている。したがって、単語となる可能性が低い音素仮説を削除することができ、照合処理量を大幅に削減することができる。さらに、前向き照合部２は、上記音素仮説を展開する際に、言語モデル格納部５および単語辞書４を参照して、上記音素仮説を構成する音素状態木の状態のうち、互いに接続可能であって上記照合に関係のある状態のみにフラグを付けるようにすることができる。したがって、その場合には、木構造化された状態のうち上記照合に関係のない状態に関するビタビ計算を行う必要がなく、照合処理量を更に削減することができるのである。

尚、上述の説明において、上記音素環境依存音響モデルは、トライフォンモデルと呼ばれる前後１つずつの音素環境を考慮したHMMを用いたが、隣接するサブワードに依存して決定されるサブワードはこれに限定されるものではない。

ところで、上記実施の形態における音響分析部１，前向き照合部２および後向き探索部８による上記音響分析手段，照合手段および検索手段としての機能は、プログラム記録媒体に記録された連続音声認識プログラムによって実現される。上記実施の形態における上記プログラム記録媒体は、ＲＡＭ(ランダム・アクセス・メモリ)とは別体に設けられたＲＯＭ(リード・オンリ・メモリ)でなるプログラムメディアである。あるいは、外部補助記憶装置に装着されて読み出されるプログラムメディアであってもよい。尚、何れの場合においても、上記プログラムメディアから連続音声認識プログラムを読み出すプログラム読み出し手段は、上記プログラムメディアに直接アクセスして読み出す構成を有していてもよいし、上記ＲＡＭに設けられたプログラム記憶エリア(図示せず)にダウンロードし、上記プログラム記憶エリアにアクセスして読み出す構成を有していてもよい。尚、上記プログラムメディアからＲＡＭの上記プログラム記憶エリアにダウンロードするためのダウンロードプログラムは、予め本体装置に格納されているものとする。

ここで、上記プログラムメディアとは、本体側と分離可能に構成され、磁気テープやカセットテープ等のテープ系、フロッピーディスク，ハードディスク等の磁気ディスクやＣＤ(コンパクトディスク)－ＲＯＭ，ＭＯ(光磁気)ディスク，ＭＤ(ミニディスク)，ＤＶＤ(ディジタル多用途ディスク)等の光ディスクのディスク系、ＩＣ(集積回路)カードや光カード等のカード系、マスクＲＯＭ，ＥＰＲＯＭ(紫外線消去型ＲＯＭ)，ＥＥＰＲＯＭ(電気的消去型ＲＯＭ)，フラッシュＲＯＭ等の半導体メモリ系を含めた、固定的にプログラムを坦持する媒体である。

また、上記実施の形態における連続音声認識装置は、モデムを備えてインターネットを含む通信ネットワークと接続可能な構成を有する場合には、上記プログラムメディアは、通信ネットワークからのダウンロード等によって流動的にプログラムを坦持する媒体であっても差し支えない。尚、その場合における上記通信ネットワークからダウンロードするためのダウンロードプログラムは、予め本体装置に格納されているものとする。あるいは、別の記録媒体からインストールされるものとする。

尚、上記記録媒体に記録されるものはプログラムのみに限定されるものではなく、データも記録することが可能である。

請 求 の 範 囲

1． 隣接するサブワードに依存して決定されるサブワードを認識単位とすると共に、サブワード環境に依存する環境依存音響モデルを用いて、連続的に発声された入力音声を認識する連続音声認識装置であって、

上記入力音声を分析して特徴パラメータの時系列を得る音響分析部（1）と、

語彙中の各単語が、サブワードのネットワークあるいはサブワードの木構造として格納された単語辞書（4）と、

単語間の接続情報を表す言語モデルが格納された言語モデル格納部（5）と、

上記環境依存音響モデルが、当該環境依存音響モデルの状態系列のうち、複数のサブワードモデルの状態系列をまとめて木構造化して成るサブワード状態木として格納されている環境依存音響モデル格納部（3）と、

上記環境依存音響モデルであるサブワード状態木，単語辞書（4）および言語モデルを参照して上記サブワードの仮説を展開すると共に、上記特徴パラメータの時系列と上記展開された仮説との照合を行い、単語の終端に該当する仮説に関する単語，累積スコアおよび始端開始フレームを含む単語情報を単語ラティスとして出力する照合部（2）と、

上記単語ラティスに対する探索を行って認識結果を生成する探索部（8）

を備えたことを特徴とする連続音声認識装置。

2． 請求項1に記載の連続音声認識装置において、

上記環境依存音響モデル格納部（3）に格納されている環境依存音響モデルは、中心サブワードが前後のサブワードに依存する環境依存音響モデルのうち、先行サブワードおよび中心サブワードが同じサブワードモデルの状態系列を木構造化したサブワード状態木であることを特徴とする連続音声認識装置。

3． 請求項2に記載の連続音声認識装置において、

上記環境依存音響モデルは、複数のサブワードモデルで状態を共有している状態共有モデルであることを特徴とする連続音声認識装置。

4． 請求項１に記載の連続音声認識装置において、

上記照合部(２)は、上記サブワード状態木を参照して仮説を展開する際に、上記単語辞書(４)および言語モデルから得られる接続可能なサブワード情報を用いて、上記仮説であるサブワード状態木を構成する状態のうち、互いに接続可能な状態にフラグを付すようになっていることを特徴とする連続音声認識装置。

5． 請求項１に記載の連続音声認識装置において、

上記照合部(２)は、上記照合を行う際に、上記特徴パラメータの時系列に基づいて上記展開された仮説のスコアを算出すると共に、このスコアの閾値あるいは仮説数を含む基準に従って上記仮説の枝刈りを行うようになっていることを特徴とする連続音声認識装置。

6． 隣接するサブワードに依存して決定されるサブワードを認識単位とすると共に、サブワード環境に依存する環境依存音響モデルを用いて、連続的に発声された入力音声を認識する連続音声認識方法であって、

音響分析部によって、上記入力音声を分析して特徴パラメータの時系列を得、

照合部によって、上記環境依存音響モデルの状態系列を木構造化して成るサブワード状態木、語彙中の各単語がサブワードのネットワークあるいはサブワードの木構造として記述された上記単語辞書、および、単語間の接続情報を表す言語モデルを参照して、上記サブワードの仮説を展開すると共に、上記特徴パラメータの時系列と上記展開された仮説との照合を行って、単語の終端に該当する仮説に関する単語,累積スコアおよび始端開始フレームを含む単語情報を単語ラティスとして生成し、

探索部によって、上記単語ラティスに対する探索を行って認識結果を生成することを特徴とする連続音声認識方法。

7． コンピュータを、請求項１に記載の音響分析部(１),単語辞書(４),言語モデル格納部(５),環境依存音響モデル格納部(３),照合部(２)および探索部(８)と

して機能させることを特徴とする連続音声認識プログラム。

8. 請求項7に記載の連続音声認識プログラムが記録されたことを特徴とするコンピュータ読出し可能なプログラム記録媒体。

Fig.1

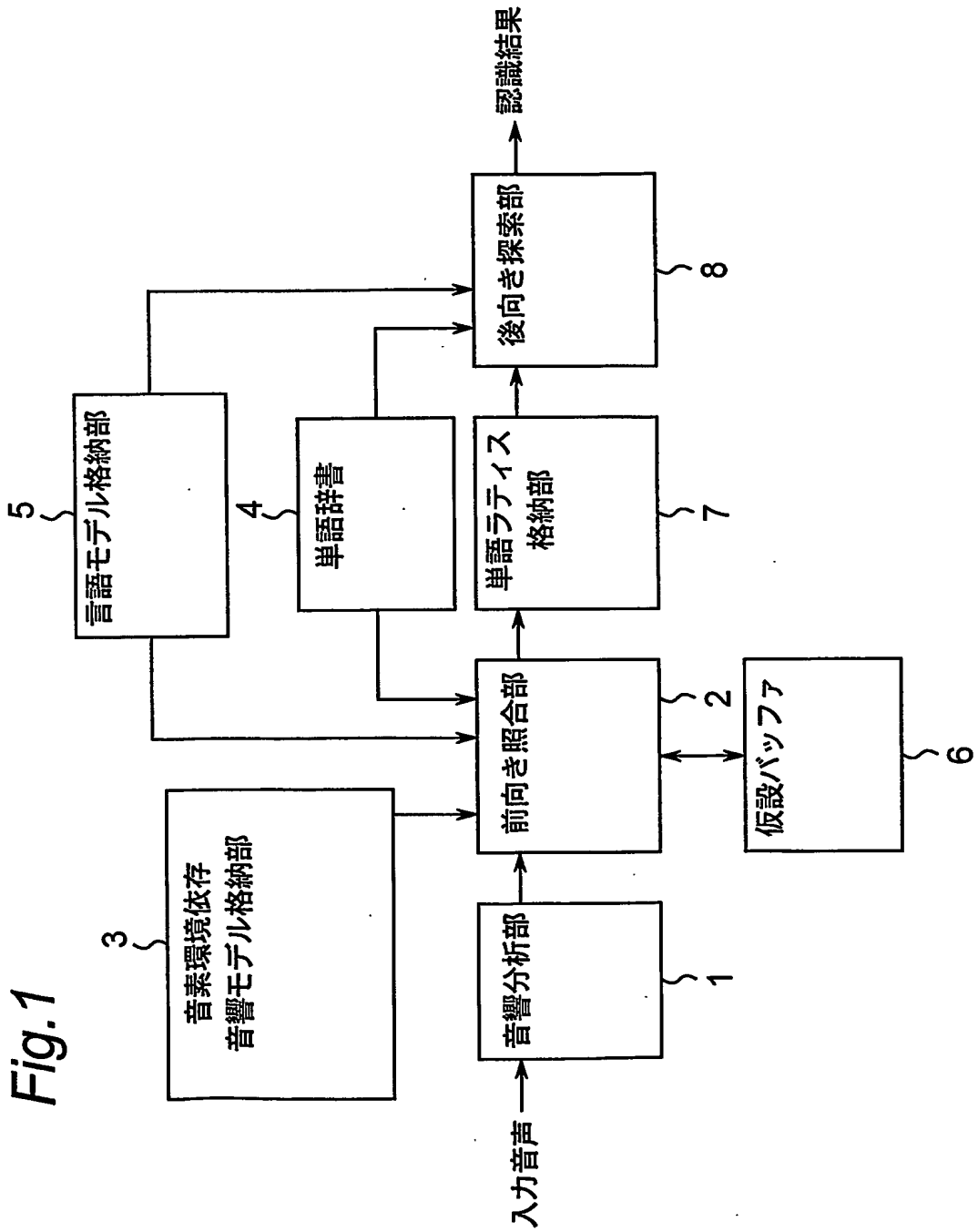

入力音声
音響分析部
1
前向き照合部
2
音素環境依存
音響モデル格納部
3
単語辞書
4
言語モデル格納部
5
仮設バッファ
6
単語ラティス
格納部
7
後向き探索部
8
認識結果

Fig.2A

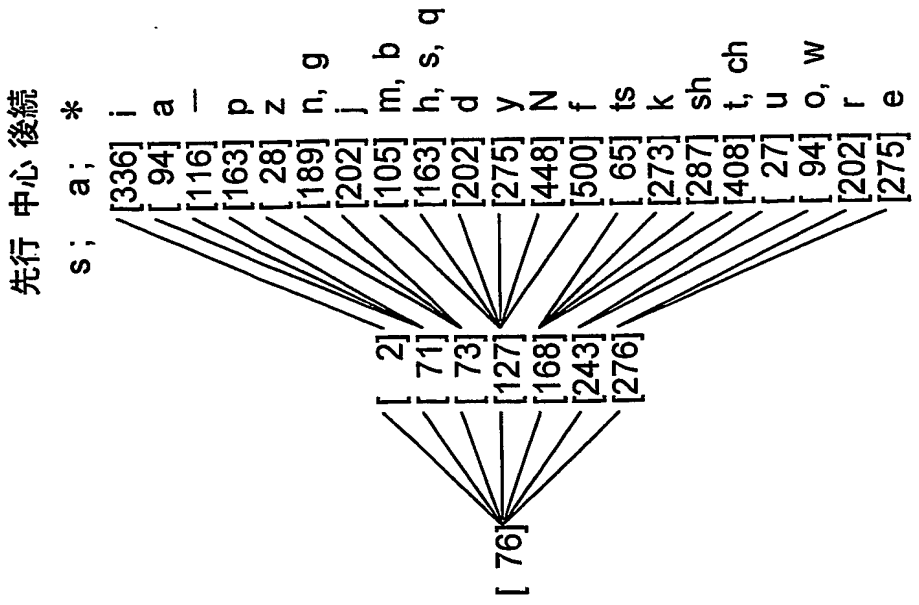
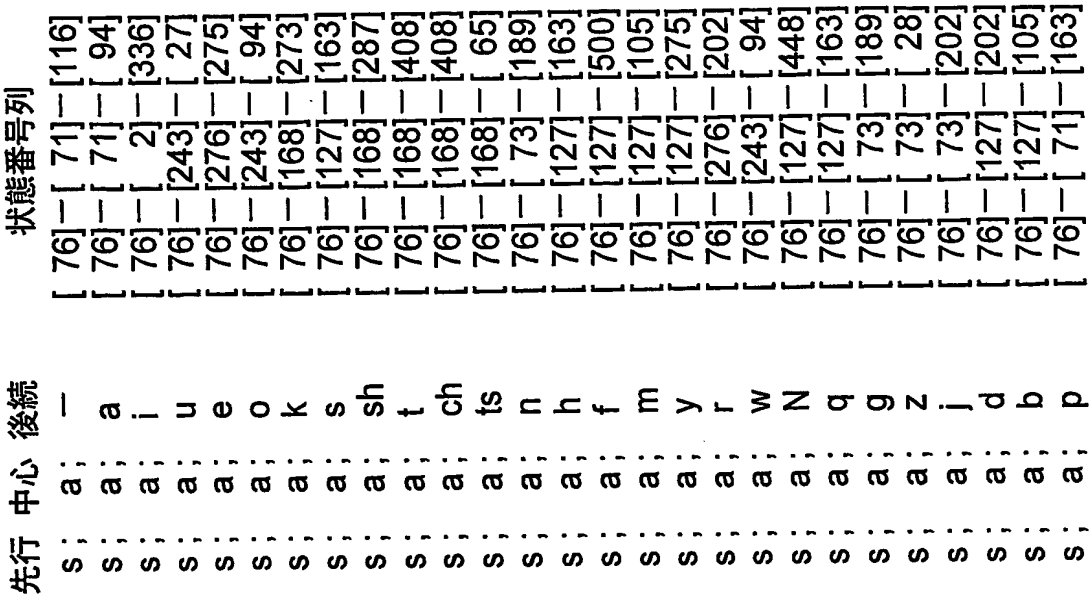

先行 中心 後続
s ; a ; *
[336] i
[94] a
[116] —
[163] p
[28] z
[189] n, g
[202] j, m, b
[105] t, h, s, q
[163] ch
[202] d
[275] y
[448] N
[500] f
[65] ts
[273] k
[287] sh
[408] t, ch
[27] u
[94] o, w
[202] r
[275] e
[76]
[2]
[71]
[73]
[127]
[168]
[243]
[276]

Fig.2B


先行 中心 後続 状態番号列
s ; a ; — [76] — [71] — [116]
s ; a ; a [76] — [71] — [94]
s ; a ; i [76] — [2] — [336]
s ; a ; u [76] — [243] — [27]
s ; a ; e [76] — [276] — [275]
s ; a ; o [76] — [243] — [94]
s ; a ; k [76] — [168] — [273]
s ; a ; s [76] — [127] — [163]
s ; a ; sh [76] — [168] — [287]
s ; a ; t [76] — [168] — [408]
s ; a ; ch [76] — [168] — [408]
s ; a ; ts [76] — [168] — [65]
s ; a ; n [76] — [73] — [189]
s ; a ; h [76] — [127] — [163]
s ; a ; f [76] — [127] — [500]
s ; a ; m [76] — [127] — [105]
s ; a ; y [76] — [127] — [275]
s ; a ; r [76] — [276] — [202]
s ; a ; w [76] — [243] — [94]
s ; a ; N [76] — [127] — [448]
s ; a ; q [76] — [127] — [163]
s ; a ; g [76] — [73] — [189]
s ; a ; z [76] — [73] — [28]
s ; a ; j [76] — [73] — [202]
s ; a ; d [76] — [127] — [202]
s ; a ; b [76] — [127] — [105]
s ; a ; p [76] — [71] — [163]

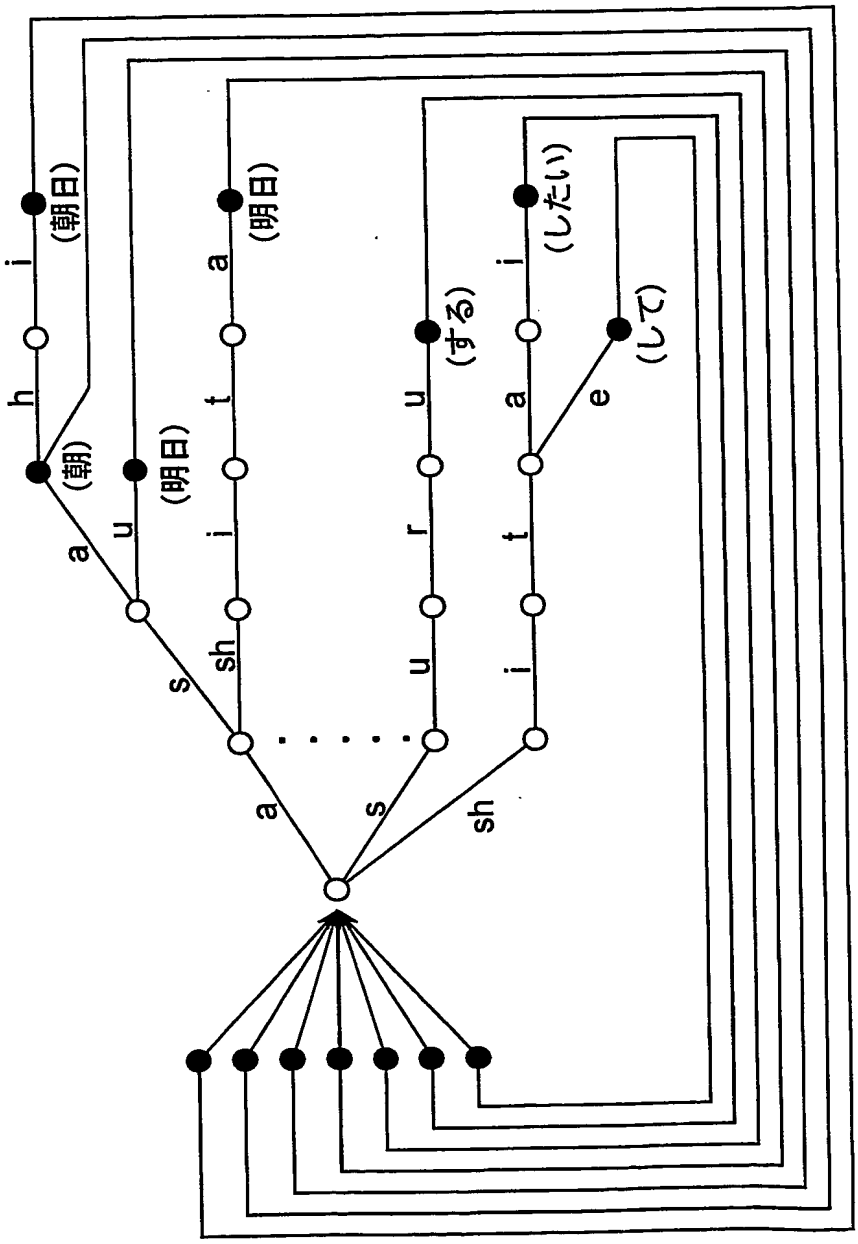


Fig.3

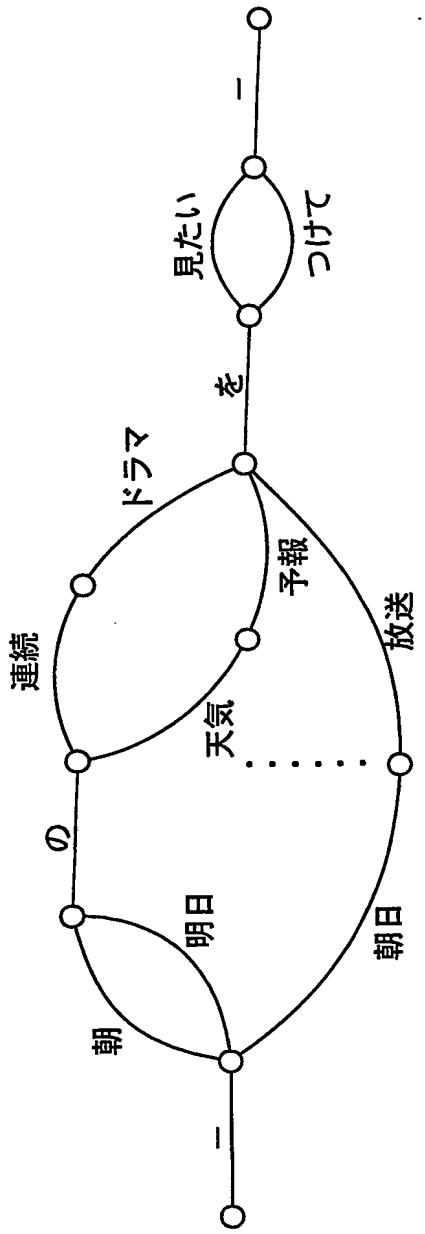


Fig.4

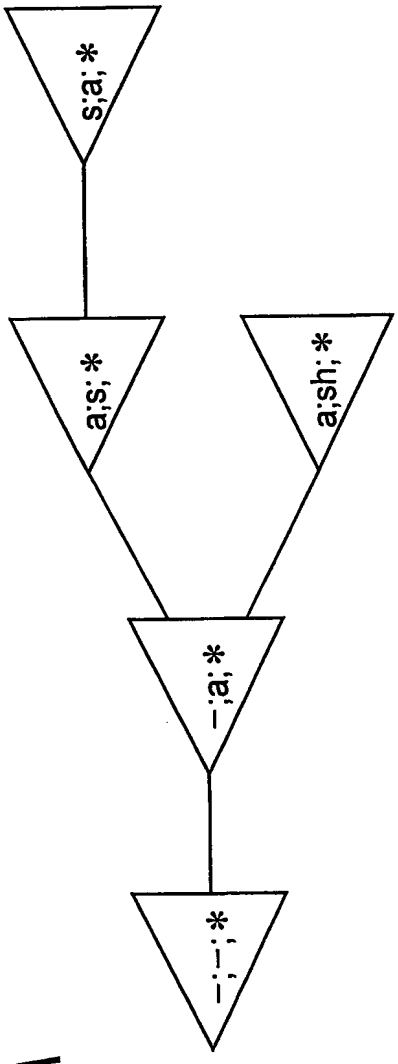

s;a;*
a;s;*
a;sh;*
-;a;*
-;-;*

Fig.5A

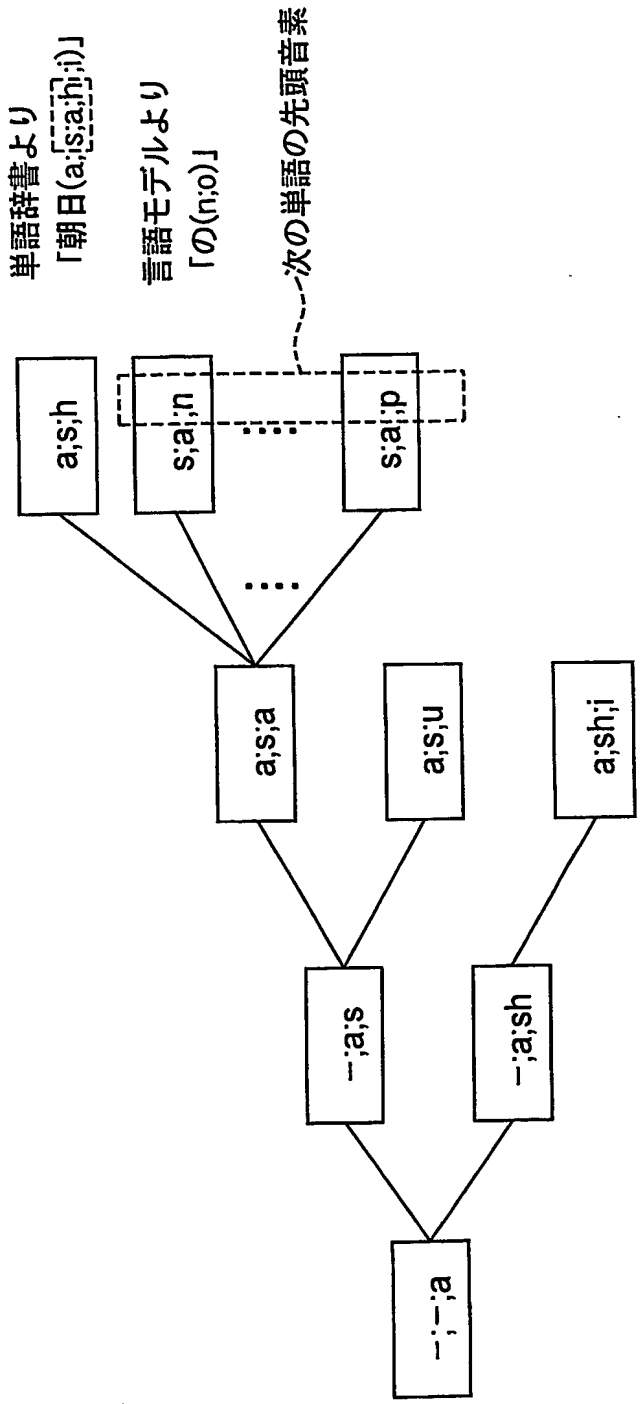

単語辞書より
「朝日(a;[s;a;h];i)」
言語モデルより
「の(n;o)」
次の単語の先頭音素
a;s;h
s;a;n
s;a;p
a;s;a
a;s;u
a;sh;i
-;a;s
-;a;sh
-;-;a

Fig.5B

Fig.6

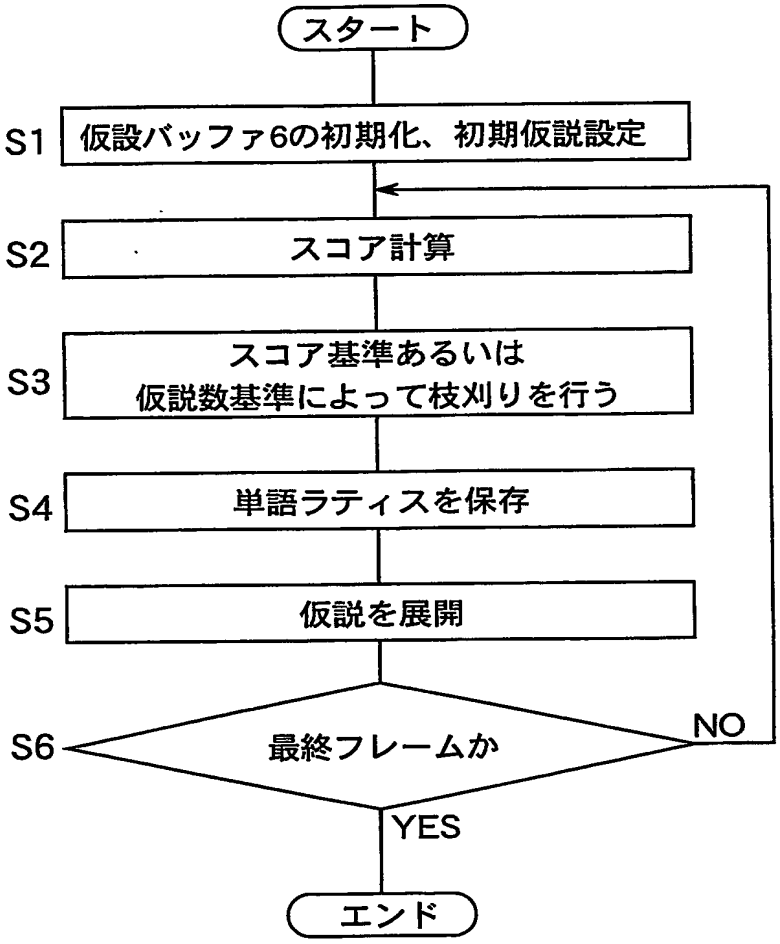

スタート
S1
仮設バッファ6の初期化、初期仮説設定
S2
スコア計算
S3
スコア基準あるいは
仮説数基準によって枝刈りを行う
S4
単語ラティスを保存
S5
仮説を展開
S6
最終フレームか
NO
YES
エンド

Fig.7A

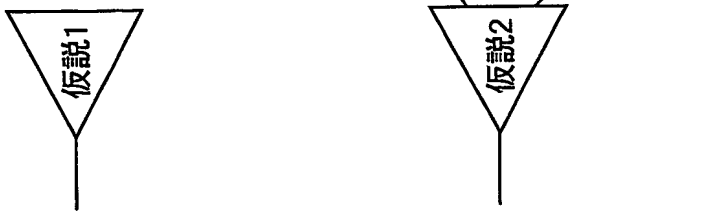


Fig.7B

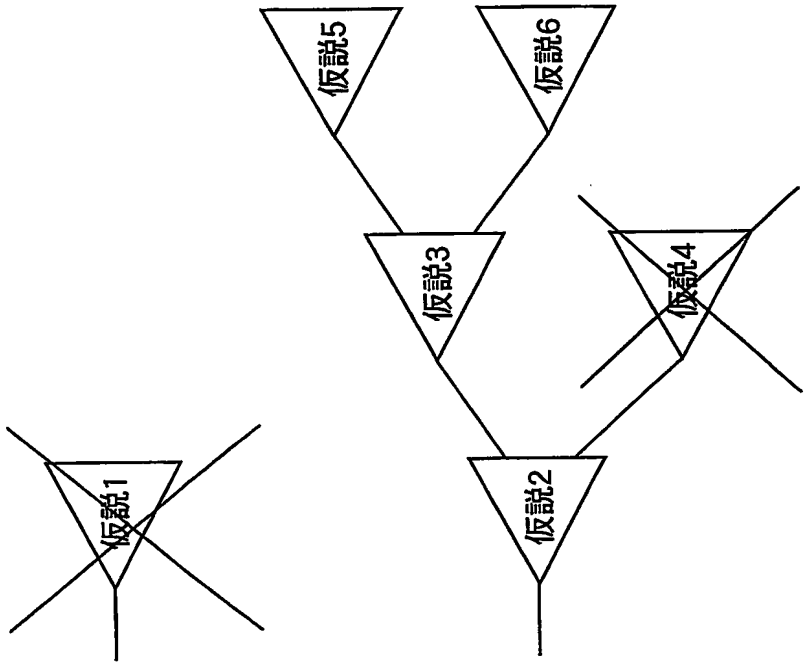


Fig.8

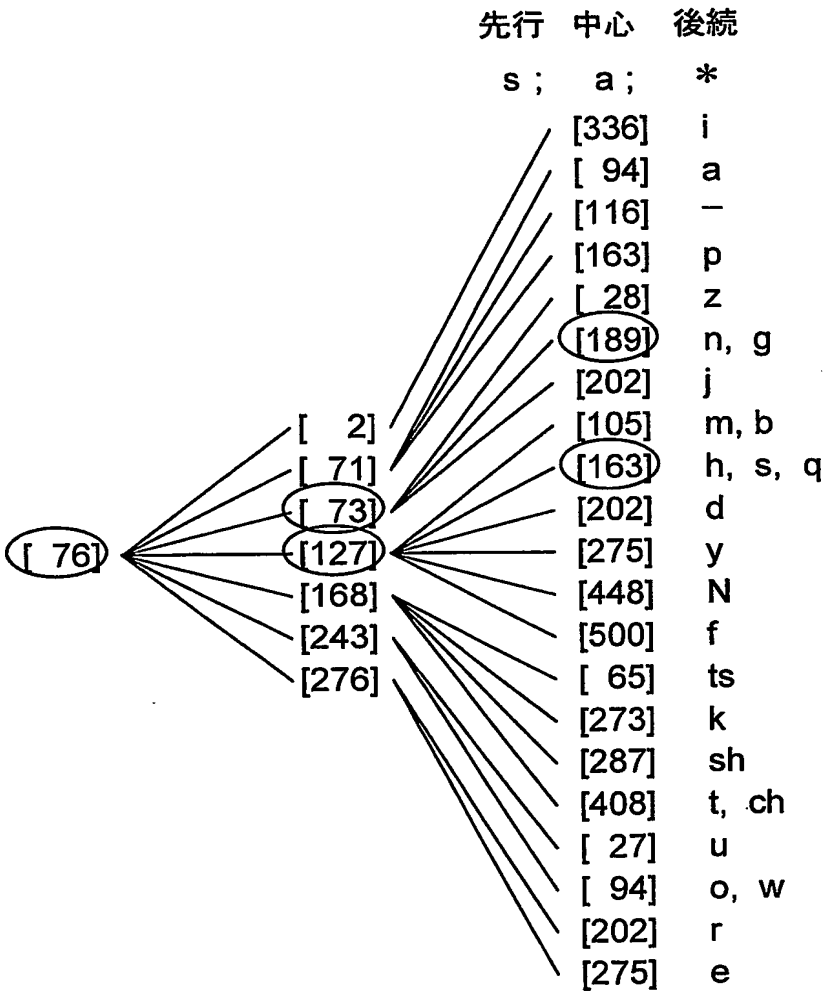

先行
中心
後続
s ;
a ;
*
[76]
[2]
[71]
[73]
[127]
[168]
[243]
[276]
[336] i
[94] a
[116] −
[163] p
[28] z
[189] n, g
[202] j
[105] m, b
[163] h, s, q
[202] d
[275] y
[448] N
[500] f
[65] ts
[273] k
[287] sh
[408] t, ch
[27] u
[94] o, w
[202] r
[275] e

Fig.9A

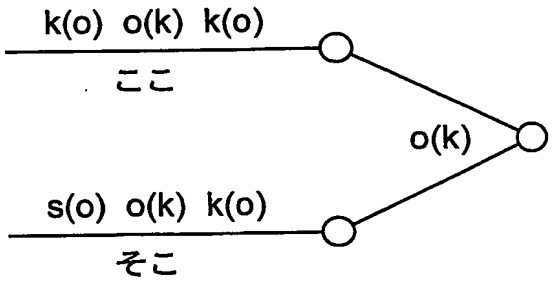
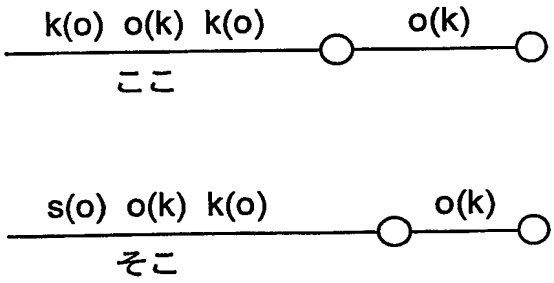


Fig.9B



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP02/13053

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[7] G10L15/06, G10L15/18

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[7] G10L15/06, G10L15/14, G10L15/18

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1926-1995	Toroku Jitsuyo Shinan Koho	1994-2003
Kokai Jitsuyo Shinan Koho	1971-2003	Jitsuyo Shinan Toroku Koho	1996-2003

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JICST FILE(JOIS), IEEE Xplore

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 97/42626 A1 (BRITISH TELECOMMUNICATION PUBLIC LTD., CO.), 13 November, 1997 (13.11.97), Full text; all drawings & JP 2000-509836 A	1-8
A	RI, KAWAHARA, TAKEDA, SHIKANO, "Phonetic Tied-Mixture Model o Mochiita Dai Goi Renzoku Onsei Ninshiki", The Institute of Electronics, Information and Communication Engineers Gijutsu Kenkyu Hokoku [Gengo Rikai to Communication], 20 December, 1999 (20.12.99), Vol.99, No.523, NLC99-32, pages 43 to 48	1-8
A	EP 1128361 A2 (SONY CORP.), 29 August, 2001 (29.08.01), Full text; all drawings & US 2001/20226 A1 & JP 2001-242884 A	1-8

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search 14 March, 2003 (14.03.03)	Date of mailing of the international search report 25 March, 2003 (25.03.03)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (July 1998)

国際調査報告 [redacted]　　国際出願番号　PC[redacted]JP02／13053

A.　発明の属する分野の分類（国際特許分類（IPC））

Int. Cl⁷　G10L15／06，　G10L15／18

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int. Cl⁷　G10L15／06，　G10L15／14，　G10L15／18

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報　　1926～1995年
日本国公開実用新案公報　1971～2003年
日本国登録実用新案公報　1994～2003年
日本国実用新案登録公報　1996～2003年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JICSTファイル（JOIS）
IEEE Xplore

C.　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	WO　97／42626　A1（BRITISH TELECOMMUNICATION PUBLIC LIMITED COMPANY）1997.11.13，全文，全図 ＆　JP2000－509836　A	1－8
A	李，河原，竹田，鹿野，「Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識」，電子情報通信学会技術研究報告［言語理解とコミュニケーション］，1999.12.20，Vol.99，No.523，NLC99-32，Pages 43-48	1－8

☒ C欄の続きにも文献が列挙されている。　　☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
14.03.03	25.03.03

国際調査機関の名称及びあて先
日本国特許庁（ISA／JP）
郵便番号100－8915
東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）　5C　9379
樫本　剛

電話番号　03－3581－1101　内線　3541

国際調査報告 [redacted]　　国際出願番号　PC[redacted]JP02／13053

C（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	EP　1128361　A2 (SONY CORPORATION) 2001.08.29, 全文，全図 & US　2001／20226　A1 & JP　2001－242884　A	1－8

様式PCT／ISA／210（第2ページの続き）（1998年7月）